图书情报工作
第 62 卷 第 13 期 2018 年 7 月

基于深度学习的数据科学招聘实体自动抽取及分析研究*

王东波[1] 胡昊天[1] 周鑫[2] 朱丹浩[3]

[1] 南京农业大学信息科学技术学院 南京 210095 [2] 南京大学信息管理学院 南京 210093

[3] 南京大学计算机科学与技术系 南京 210093

摘要：［目的/意义］数据科学作为一个融合诸多领域的新兴交叉学科正在快速形成。从数据科学招聘的公告信息中，抽取出相应的实体知识不仅有助于从市场的角度了解数据科学的发展动态，而且有助于改进数据科学教学的内容。［方法/过程］基于各大招聘网站职位招聘公告，结合情报学的数据获取、标注和组织方法，构建数据科学招聘语料库并从中抽取相应的实体进行分析与研究。［结果/结论］在搜集到的 11 000 篇经过标注的职位招聘公告语料的基础上，基于 Bi-LSTM-CRF、CRF 和 Bi-LSTM 模型，对数据科学招聘实体的抽取任务进行性能的对比，确定最终的数据科学招聘实体自动抽取模型，设计数据科学招聘实体自动抽取平台，并构建数据科学招聘实体网络。

关键词：数据科学 条件随机场 深度学习 Bi-LSTM-CRF

分类号：G255.1

DOI：10.13266/j.issn.0252-3116.2018.13.009

1 引言

数据科学作为一个融合了计算机科学、统计学、应用与计算数学、人工智能、系统科学、社会科学、心理学、经济学等诸多领域的新兴交叉学科正在快速崛起。在这一发展趋势下，与数据科学相关的数据科学家、数据分析师、数据标注师和数据工程师等职位大量涌现。获取与数据科学相关的招聘职位要求，并从非结构化的招聘信息中，通过机器学习的策略从中抽取职位名称、要求专业、学历要求、经验要求、能力水平、所掌握的编程语言和算法等实体，对此进行分析与研究一方面有助于数据科学工作者掌握市场对数据科学人才的具体需要，从而有针对性地提升自身应对数据科学工作的能力；另一方面有益于数据科学的教育者拟定数据科学的教育体系和人才培养的目标。

目前，对于英文实体的抽取，国外的相关研究工作取得了较为满意的效果。英文实体主要涵盖了人名、地名、机构名、时间、数字、货币等类别，而英语实体抽取主要是通过基于规则、统计和机器学习的策略完成对上述几类实体识别的任务。M. M. Bikel 等[1]基于隐马尔科夫模型设计了抽取人名、地名和组织名的方法且取得了很好的效果。相较于基于规则和简单的统计方法来说，隐马尔科夫模型能够充分利用实体右边界的特征，该模型的这一特征确保了实体识别的精准率。A. L. Berger 等[2]基于最大熵模型提出了实际使用效果较为优秀的抽取方法。由于在构建最大熵实体识别模型的过程中可以选择跟序列有关的特征，从而确保了基于最大熵构建的实体模型无论是在精准率还是召回率上均优于隐马尔科夫模型。J. Lafferty 等提出了条件随机场模型[3]，这一模型结合了隐马尔科夫模型与最大熵模型的优点，确保了在实体这一识别任务上的性能是最为突出的。M. C. Callum 等[4]将 CRF 模型应用到实体自动抽取中且验证了 CRF 模型的识别效果要优于隐马尔科夫模型和最大熵模型。因为条件随机场模型不仅能够利用实体左右边界的特征，而且可以把任何有益于实体识别的特征融入到条件随机场模型当中，从而确保了所构建的实体识别模型的整体性能较为突出。中文句子与英文句子在词法与句法上的差异，使得中文实体抽取的难度更大。加之国内

* 本文系国家社会科学基金重大项目“情报学学科建设与情报工作未来发展路径研究”（项目编号：17ZDA291）和江苏省普通高校学术学位研究生科研创新计划项目“引用内容分析--引文语义信息的自动挖掘（KYZZ16_0033）”研究成果之一。

作者简介：王东波（ORCID：0000-0002-9894-9550），副教授，硕士生导师，E-mail：db.wang@njau.edu.cn；胡昊天（ORCID：），本科生；周鑫（ORCID：），博士研究生；朱丹浩（ORCID：），助理馆员。

收稿日期：2017-12-02 修回日期：2018-04-08 本文起止页码：64-73 本文责任编辑：易飞

对实体抽取的研究起步较晚,因此中文实体抽取的研究较英文相对落后。张小衡等[5]基于人工规则提出了一种抽取高校名称的方法。这一研究是典型的基于规则策略下的实体识别探究,虽然性能整体不高,但对于高校名称的规则分布进行了比较细致的分析。Y. Zhang 等[6]基于记忆的学习算法,开发出识别命名实体及其之间关系的系统。从统计学的角度,这一研究有机地利用了实体左右边界的特征,从研究方法上看具有一定的借鉴价值和意义。郑逢强等[7]将义原作为特征加入到最大熵模型中以提高其抽取性能。这一研究不仅发挥了最大熵可以利用实体边界特征的属性,而且把语义知识融入到模型的构建当中,从领域知识利用的角度分析,这一研究具有一定的创新性。陈宇等[8]尝试用基于神经网络的方法对实体及实体之间的关系进行抽取。虽然神经网络能够充分挖掘实体的特征,但这一研究的识别效果有很大的提升空间。邵发等[9]利用消除歧义的方法,通过利用 HowNet 和贝叶斯分类抽取实体,从而解决一词多义的问题,从方法论的角度来看,这一研究把实体识别的任务转化成了分类的问题,并且把深层次的语义知识融入到了分类模型当中,具有较强的创新性。许华等[10]基于分词、词性标注的医疗语料,利用规则的方法,完成了对医疗文本中实体的抽取且整体性能较高。这一研究从所使用的方法上看没有太大的创新性,但对医疗这一领域化的实体进行识别具有领域知识挖掘上的创新性。基于深度学习进行实体抽取的研究是近两年兴起的探究方法,比较有代表性的研究如下:冯蕴天和张宏军等[11]在前人研究的基础上利用深度信念网络对神经网络语言模型进行了扩展,提出了一种可用于命名实体识别的深层架构。这一结构框架对于实体的识别不仅具有宏观上的指导性,而且具有方法论上的引导性。C. Dong 和 J. Zhang 等[12]首次将基于字符级的 BILSTM-CRF 神经结构用于中文命名实体识别,在第三届 SIGHAN Bakeoff MSRA 数据集上取得不错的效果。这一研究验证了 BILSTM-CRF 组合的优势,并且为基于字的汉语实体的识别奠定了坚实的基础。朱丹浩和杨蕾等[13]基于 RNN 方法,重新定义了机构名标注的输入和输出,提出了汉字级别的循环网络标注模型。这一研究首次把深度学习应用在了机构实体的识别上,具有方法论上的借鉴意义和价值。

针对中国的具体情况,结合数据科学的发展,国内的相关研究者对数据科学的研究情况进行了多个角度的探究,具体如下:叶鹰和马费成[14]指出了数据科学与信息科学在理论逻辑和技术方法上一脉相承,揭示了数据科学继续维持信息科学基本原理。这一研究从理论的角度探究了数据科学与信息科学的关系,为数据科学的发展提供了坚实的理论支撑。杨京和王效岳等[15]分析了大数据给数据科学分析工具带来的挑战,介绍了应运而生的大数据分析工具及其发展趋势。这一研究从大数据的角度对数据科学分析工具的开发指明了方向。周傲英和钱卫宁等[16]论述了数据科学与工程这一新兴交叉学科的发展必然性,阐述了其学科特点、知识体系和建设思路。以工程为切入点,这一研究丰富了数据科学的内涵和外延。朝乐门和卢小宾[17]提出数据科学将成为信息科学领域知识的新理论基础,并指出了大数据时代信息科学研究的新课题。这一研究把数据科学放到了信息科学这一大的框架下,厘清了数据科学和信息科学的关系。王曰芬和谢清楠等[18]利用 Web of Science 核心合集数据库从数据科学的内涵界定与应用方向对国外有关数据科学的文献进行计量分析,基于文献计量学,这一研究系统而全面地总结了数据科学国内外的研究情况,为了解和掌握数据科学的发展趋势提供了第一手的资料, 为我国今后的研究提供了参考与借鉴。

在上述相关研究的基础上,面向国内的主要招聘网站,通过设定与数据科学相关的关键词,抓取 29 460 篇职位招聘公告,并在人工标注 11 000 篇数据科学招聘公告实体的基础上,构建中国数据科学招聘语料库。基于该语料库,通过测试条件随机场和深度学习的相关模型,构建面向数据科学招聘公告的实体自动抽取模型,并搭建相应的平台,更进一步地基于复杂网络对相关的实体分布情况进行分析。

2 数据科学招聘语料简介及实体界定

在对智联招聘、51job 等网站上面有关数据科学的招聘职位数据进行抓取、清洗、标注和组织的基础上,本文构建了数据科学职位招聘语料库,具体流程如下:

(1)数据科学职位招聘数据的采集内容主要来自招聘网站上有关数据科学的职位信息。基于 2017 年 3 月至 2017 年 8 月间的招聘信息,本文利用 Python 开发的网络爬虫工具抓取了 29 460 条职位招聘公告。

(2)基于所抓取的数据,选取招聘职位信息这一个字段,提取出 29 460 条招聘职位的描述信息。一方面通过去重算法,完成对 29 460 条招聘信息的自动去重,共获得 24 460 条去重后的招聘信息;另一方面,在招聘描述信息里面,有一些为英语招聘信息,由于本文

主要所涉及到的是汉语招聘的信息，对这一部分英文的招聘信息也进行了清理，最后得到 23 154 条数据科学招聘信息。

（3）本文所谓的数据科学招聘实体主要是指招聘详细信息语料中涉及到的职位名称、学历要求、经验要求、能力水平和相关软件等实体内容，比如具体的职位名称有“软件工程师、数据分析师、数据库工程师”等，具体学历要求为“本科及以上学历、大专及以上学历”等，具体相关软件为“MYSQL、Python、Java、Spark、SAS、SPSS、R”等。数据科学招聘实体这一概念是借鉴实体这一概念的内涵和外延，结合数据科学招聘这一特定领域而确定的。在确定的上述 5 类数据科学实体基础上，制定相应的标注规则，由 55 名标注人员完成了对其中 11 000 篇数据科学招聘职位文本内容实体的标注。具体标注后的样例如图 1 所示：

【大学本科或以上】学历，【数理统计】、【信息技术】、【信息系统】等专业优先考虑；2、【一年以上】【数据挖掘】等相关经验；3、能够【独立工作】，【执行力】强；具备良好的【团队协作能力】；具备良好的【沟通能力】；4、掌握【统计分析】、【分类】、【聚类】、【回归】、【关联规则】和【时间序列】等【数据分析】和【挖掘】方法；5、掌握【SAS】、【SPSS】、【R】、【Python】等至少一种数据分析软件；6、熟悉【数据库】和【SQL】尤佳。工作地址：深圳市龙华区汇海广场B座15楼

图 1　数据科学招聘实体标注样例

3　机器学习模型简介

在本节中，本文对条件随机场（CRF）模型、长期短期记忆网络（LSTM）模型和 LSTM-CRF 模型进行了简要介绍。

3.1　CRF 模型

条件随机场是用于解决序列标注问题较新的一种模型，是指在给定一组需要标记的观察序列的条件下，计算整个观察序列状态标记的联合条件概率分布的无向图模型，其拓扑结构如图 2 所示：

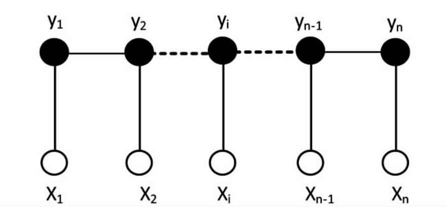


图 2　线性链 CRFs 模型的拓扑结构

设 $x=\{x_1,x_2,\cdots,x_{n-1},x_n\}$ 表示被观察到的输入数据序列，$y=\{y_1,y_2,\cdots,y_{n-1},y_n\}$ 表示有限状态集合，其中每个状态对应于一个标记。在给定输入序列 x 的条件下，对于参数 $\lambda=\{\lambda_1,\lambda_2,\cdots\lambda_{n-1},\lambda_n\}$ 的线性链 CRFs 的状态序列 y 的条件概率为：

$$p(y|x,\lambda)=\frac{1}{z_x}\exp(\sum_{i=1}^{n}\sum_j\lambda_j f_j(y_{i-1},y_i,x,i))$$ 式(1)

$$z_x=\sum_y\exp(\sum_{i=1}^{n}\sum_j\lambda_j f_j(y_{i-1},y_i,x,i))$$ 式(2)

其中 z_x 为归一化因子，表示所有可能的状态序列的得分，确保所有可能状态序列的条件概率之和为 1。$f_j(y_{i-1},y_i,x,i)$ 是一个统一形式的特征函数，通常为二值表征函数；λ_j 是通过模型对训练数据进行训练之和获得的相应特征函数的权重。在构建数据科学实体识别模型的过程中，这一模型不仅可以利用实体的左边界特征，而且可以利用实体的右边界特征，从而确保了所构建模型的整体性能要优于隐马尔科夫模型和最大熵模型。

3.2　LSTM 模型

循环神经网络（recurrent neural network，RNN）针对前馈神经网络处理连续的序列输入没有反馈机制的问题，对各个隐藏层进行了关联。将输入集 $\{x_0,x_1,\cdots,x_t,x_{t+1},\cdots\}$ 作为一个输入向量序列并返回另一个向量序列输出集 $\{y_0,y_1,\cdots,y_t,y_{t+1},\cdots\}$。在 t 时刻时，RNN 隐藏层和输出层的计算公式为：

$$h_t=f(Ux_t+Wh_{t-1})$$ 式（3）

$$y_t=g(Vh_t)$$ 式（4）

在公式（3）和（4）中，x 为输入层，h 为隐藏层，y 为输出层。U、W 和 V 分别是 RNN 中输入层到隐藏层、前后两个隐藏层之间及隐藏层到输出层的权值，f 和 g 是非线性激活函数 sigmoid 和 softmax 激活函数。虽然在理论上 RNN 可以学习长期的依赖关系，但实际效果并不良好，长期短期记忆网络（long short-term memory，LSTM）正是为了解决这一问题而提出的。LSTM 通过结合一个记忆单元（memory cell），并引入门（gate）控制器来控制历史信息的保留和丢弃。LSTM 记忆单元的计算公式如下：

$$i_t=\sigma(W_i h_{t-1}+U_i x_t+b_i)$$ 式(5)

$$f_t=\sigma(W_f h_{t-1}+U_f x_t+b_f)$$ 式（6）

$$o_t=\sigma(W_o h_{t-1}+U_o x_t+b_o)$$ 式(7)

$$c_t=f_t\odot c_{t-1}+i_t\odot\tanh(W_c h_{t-1}+U_c x_t+b_c)$$ 式(8)

$$h_t=o_t\odot\tanh(c_t)$$ 式(9)

其中的激活函数 σ 一般选取 sigmoid 函数，⊙是表示点乘运算。公式（5）、（6）和（7）中的 i_t，f_t 和 o_t 分别表示的是 t 时刻的输入控制门、遗忘控制门和输出控制门。公式（8）中的 c_t 表示的是 t 时刻记忆单元向量。U_i，U_f，U_c，U_o 分别是输入序列 $\{x_0,x_1,\cdots,x_t,x_{t+1},\cdots\}$ 和各个控制门之间的连接权重矩阵，并且是控制门和

隐藏状态 h 之间的连接权重矩阵。由于本研究是按字为最小单位进行的数据科学命名实体识别，在模型训练过程中，不仅需要考虑当前字与前文的联系，还要结合使用后文的信息进行预测和序列标注任务。而双向 LSTM(Bi-LSTM)拥有两个相反方向的并行层，能够存储来自两个方向的信息。因此，本文选择双向 LSTM (Bi-LSTM)来处理实体标注的任务。

3.3 LSTM-CRF 模型

尽管通过 LSTM 网络可以获得较好的实体标注效果，但是当输出标签之间存在较强的依赖关系时，LSTM 模型的性能将会受到影响。特别是在实际的序列标注任务时，由于神经网络结构对数据的依赖很大，数据量的大小和质量也会严重影响模型训练的效果。为了解决这个问题，本研究采用了 LSTM-CRF 模型。LSTM-CRF 模型不仅保留了 LSTM 能够同时考虑数据科学实体的上下文信息的特性，还能够通过 CRF 层考虑输出独立标签之间前后的依赖关系，图 3 所示是用于实体识别的 LSTM-CRF 模型结构：

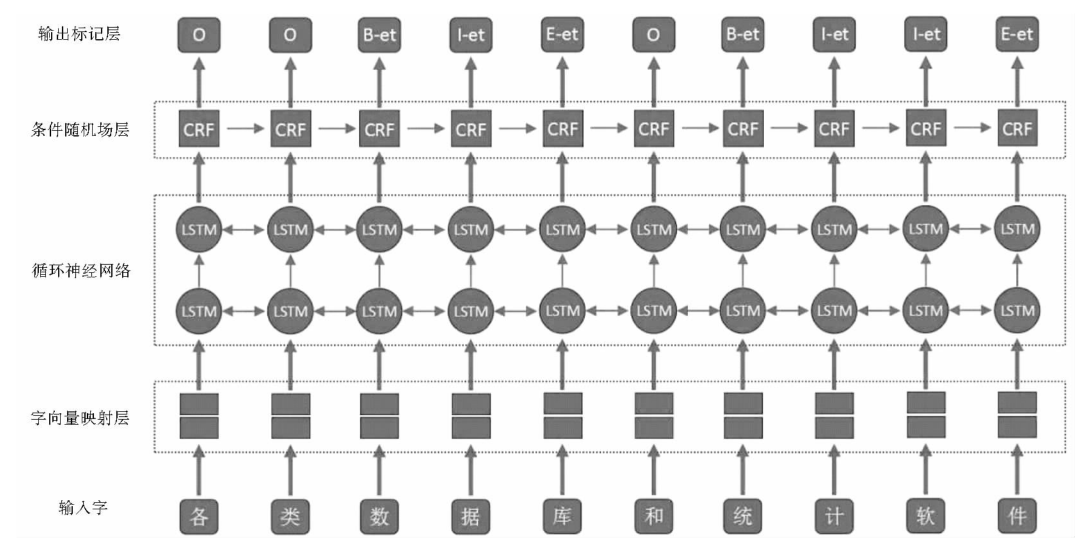


图 3 LSTM-CRF 模型的主要架构

在 LSTM-CRF 模型下，输出的将不再是相互独立的标签，而是最佳的标签序列。对于输入：$X = \{x_1, x_2, \cdots, x_n\}$，我们可以定义 A 为状态转移矩阵，P 为 LSTM 输出的概率矩阵。其中 $A_{i,j}$ 表示时序上从第 i 个状态转移到第 j 个状态的概率，$P_{i,j}$ 指观察输入序列中的第 i 个数据科学实体字被标记为第 j 个标签的概率。通过求得最大的 s(X,y)，即可得到最佳的输出标签序列，然后使用动态规划算法进行计算，得出最优路径并进行标注。对于待预测的标签序列 $y = \{y_1, y_2, \cdots, y_n\}$ 的预测输出计算公式为：

$$s(X,y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \qquad \text{式(10)}$$

4 实体识别实验

4.1 语料的预处理

基于人工标注的数据科学招聘职位中的实体真实的长度情况的描述，本文确定在不同的模型当中使用 4 字位的标注集，标注集用 R 来表示，具体为 R = {B-et, I-et, E-et, O}，B-et 表示数据科学实体的初始字，I-et 为数据科学实体的中间字，E-et 为数据科学实体的结束字，O 表示数据科学实体外字，如果数据科学实体的长度超过了 3，就用 I-et 表示扩展字。本文通过编写 Python 程序，结合语料中数据科学实体的“【】”标记，自动对所有语料进行基于字的训练和测试的标注。

由于在基于深度学习训练实体识别过程中需要使用到 GPU，因此对本文的实验环境介绍如下：CPU：Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz 内存：16GB DDR4；GPU：NVIDIA Quadro K1200 显存：4GB GDDR5；操作系统：ubuntu 16.04。服务器上高性能的 GPU 可以支持大规模的并行运算。

4.2 实体识别判定标准

本文对数据科学实体识别模型性能的评价主要采用 3 个指标来衡量：准确率(precision)、召回率(recall)、F 值(F-measure)。具体计算公式如下：

$$准确率：P = \frac{A}{A+B} \times 100\% \qquad \text{式(11)}$$

$$召回率：P = \frac{A}{A+C} \times 100\% \qquad \text{式(12)}$$

$$调和平均值：F = \frac{2 \times P \times R}{P+R} \times 100\% \qquad \text{式(13)}$$

其中,A 表示正确识别数据科学实体个数,B 表示错误识别数据科学实体个数,C 表示未识别出来的数据科学实体个数。需要说明的是,正确率(accuracy)的高低无法准确反映模型的好坏,于是本文没有使用此评价指标。

4.3 实体识别的效果分析

本文基于人工标注的 11 000 篇数据科学招聘语料使用 CRF、Bi-LSTM、和 Bi-LSTM-CRF 模型进行数据科学实体的识别。在具体的实验中使用十次交叉验证的方法来测试所构建模型的性能,将 11 000 篇语料文档分别按照 9:1 的比例分为训练语料和测试语料进行实验。测试结果如表 1 – 表 3 所示:

表 1 基于 CRF 模型的数据科学实体识别性能比较

测试编号	准确率(%)	召回率(%)	F 值(%)
1	86.21%	85.25%	85.73%
2	85.94%	85.58%	85.76%
3	85.89%	85.45%	85.67%
4	85.72%	86.06%	85.89%
5	86.25%	85.65%	85.95%
6	85.76%	85.50%	85.63%
7	85.18%	85.62%	85.40%
8	85.80%	85.92%	85.86%
9	85.18%	86.37%	85.77%
10	85.36%	85.59%	85.47%
均值	85.73%	85.70%	85.71%

从表 1 可以看出,基于条件随机场模型,所构建的以字为单位的数据科学实体识别模型的 F 平均值达到了 85.71%。这一 F 值从一个侧面说明了条件随机场模型能够充分利用数据科学实体的左右边界字的特征并这把这一特征融入到模型的构建当中。从具体识别出来的实体结果来看,专业名称、软件和模型等名称识别的整体效果较好,但对于边界界定容易模糊的实体则容易识别错误或者没有识别,比如"对专业【数据的分析】及做好竞争对手【数据的采集】、统计、评估与分析,并【编制报表】"这一识别任务中,本来要识别的实体为"分析""采集""统计""评估""分析"等表示能力的实体,但由于"数据"的出现频次过高,被作为特征概率融入到了条件随机场模型当中,造成了左边界识别的错误。

从表 2 中可以看出,由于双向 LSTM(Bi-LSTM)拥有两个相反方向的并行层特征,这一特征确保了数据科学实体识别的精准率。与 CRF 所构建的模型进行对比可以看出,基于 Bi-LSTM 所构建的数据科学实体模型在精准率上平均提升了 1.43%,在一定程度上表明了这一模型的性能要优于 CRF。在具体识别的例子上,对于"对专业数据【分析】及做好竞争对手数据的【采集】、统计、评估与【分析】,并【编制报表】"这一表述中的实体的识别,就精准地把"数据"与"分析"和"采集"进行了分割,但在这一表述中,对于"统计"和"评估"这两个实体还是未能识别出来。

表 2 基于 Bi-LSTML 模型的数据科学实体识别性能比较

测试编号	准确率(%)	召回率(%)	F 值(%)
1	87.21%	85.36%	86.28%
2	86.93%	86.68%	86.80%
3	85.99%	87.85%	86.91%
4	87.77%	88.09%	87.93%
5	87.26%	88.67%	87.96%
6	87.77%	86.59%	87.18%
7	87.19%	87.64%	87.41%
8	87.89%	86.93%	87.41%
9	86.19%	87.38%	86.78%
10	87.39%	85.79%	86.58%
均值	87.16%	87.10%	87.13%

表 3 基于 Bi-LSTM-CRF 模型的数据科学实体识别性能比较

测试编号	准确率(%)	召回率(%)	F 值(%)
1	91.35%	90.80%	91.07%
2	91.31%	90.99%	91.15%
3	90.92%	91.69%	91.31%
4	91.18%	91.79%	91.49%
5	91.16%	91.25%	91.21%
6	90.91%	91.33%	91.12%
7	90.63%	90.30%	90.47%
8	91.47%	90.98%	91.22%
9	91.20%	91.24%	91.22%
10	90.21%	91.38%	90.79%
均值	91.03%	91.18%	91.10%

从表 3 可以看出,基于 Bi-LSTM-CRF 模型的数据科学实体识别性能整体较为良好,各组的识别准确率和召回率均超过了 90%,从一定程度上充分反映出了这一组合的模型不仅保留了 LSTM 能够同时考虑上下文信息的特性,还能够通过 CRF 层考虑输出独立标签之间前后的依赖关系,从而切实地确保了数据科学识别模型的精准率和召回率。具体识别的例子体现如下,从"对专业数据【分析】及做好竞争对手数据的【采集】、【统计】、【评估】与【分析】,并【编制报表】"这一表述的识别结果来看,这一模型不仅精准地把"数据"与"分析"和"采集"进行了切分,而且对"统计"和"评估"这两个 CRF 和 Bi-LSTM 模型没有识别出来的实体

精准地进行了识别。Bi-LSTM-CRF 模型的 F 值最低为 90.47%，最高达到 91.49%，平均 F 值为 91.10%。从整体上优于 Bi-LSTM，其 F 值高于 Bi-LSTM 3.97%，这从一定程度上说明了在融入 CRF 模型获取的特征基础上，确实能够有效地提高整个序列化模型的性能。相比 CRF，Bi-LSTM-CRF 的平均 F 值高出了 5.39%，这充分说明，在字这一层级上，深度学习模型能够充分发挥端到端的模型训练和大量语料的场景特征。深度学习模型的性能从 Bi-LSTM 所构建的模型性能平均比 CRF 高出 1.42% 也直接说明了其自身的优越性。总之，仅仅基于字这一汉语的基本构成元素，在无任何人为特征添加的情况下，所构建的 Bi-LSTM-CRF 实体识别模型达到了可以应用的水平，这一探究对于其他类似序列化实体识别的研究任务具有一定程度上的借鉴意义。

4.4 搭建面向数据科学招聘的实体自动抽取平台

数据科学招聘实体自动抽取实验涉及步骤较为复杂，如数据科学招聘实体语料需要生成 Bi-LSTM-CRF 可识别的以整行形式存在的 tokens 并制作相应的特征模板，在对语料进行训练和测试后，还需要计算出其精确率 P、召回率 R 以及调和平均值 F 这 3 个评价指标。为了便于实验操作，帮助读者理解，本文调用基于 Bi-LSTM-CRF 构建最优数据科学招聘实体自动抽取模型，针对实验设计了可视化操作系统，并在此基础上构建了数据科学招聘实体自动抽取平台。

数据科学招聘实体自动抽取平台使用 Python 语言的第三方工具包 PyQt 进行开发。PyQt 是由 P. Thompson 开发的 Python 语言的图形用户界面(GUI)编程解决方案，它是 Python 编程语言和 Qt 库的成功融合。PyQt 实现了一个 Python 模块集。它有超过 300 类，将近 6 000 个函数和方法。它是一个多平台的工具包，可以运行在所有主要操作系统上，包括 UNIX，Windows 和 Mac。相对于 wxPython、Tkinter 等图形库，PyQt 功能强大，可以使用“Designer”或“Qt Creator”很方便地设计 UI 文件，从而简化了 UI 的设计布局等工作。

该平台主要由两部分组成，第一部分是数据采集与清洗功能，包括网页爬虫与脏数据清洗；第二部分是实体抽取与统计功能，包括选择语料库、抽取实体与统计词频。

使用数据采集与清洗功能时，首先点击下拉框控件选择所需招聘公告的发布时间范围，可以选择的有：24 小时内、近 3 天、近 1 周、近 1 月和全部时间。时间范围选取完毕后，点击“获取数据”按钮，平台自动启用网页爬虫抓取招聘网站上的相关职位招聘公告，并在提示框内显示抓取进度，见图 4。公告抓取完毕后，平台自动清洗数据，并将全部语料保存在指定路径。

使用实体抽取与统计功能时，点击“浏览”按钮，即可在文件夹浏览视图中选择语料库(语料库根目录)，系统自动读取语料库内全部文档路径。点击“抽取实体”按钮后，平台对语料库内全部语料进行预处理并按 Bi-LSTM-CRF 可识别的以整行形式存在的 tokens 格式要求生成“test”命名的文本文档，继而自动调用 windows 环境下命令提示符(cmd)程序，调用数据科学招聘实体自动抽取模型对 test 文档进行数据科学招聘实体自动抽取，并在“信息提示框”内显示所抽取全部数据科学招聘实体，见图 5。点击“统计词频”按钮后，平台对基于数据科学招聘实体自动抽取模型抽取的实体进行频次统计，并在“信息提示框”内按降序排列显示数据科学实体频次，见图 6。



图 4　数据科学实体自动抽取平台数据采集功能截图



图 5　数据科学实体自动抽取平台抽取实体功能截图



图 6 数据科学实体自动抽取平台统计词频功能截图

5 数据科学招聘实体的网络分析

在基于 11 000 篇数据科学招聘语料库所构建的 Bi-LSTM-CRF 抽取模型基础上，通过数据科学招聘的实体自动抽取平台，完成了对 12 154 篇通过网络爬虫所抓取的数据科学招聘新语料中实体的抽取，经过人工辅助校对，形成了 23 154 篇数据科学招聘实体的抽取。根据数据科学招聘实体分布情况以及实体的共现，本文发现数据科学招聘实体之间存在一定的连通性，一定规模的数据科学招聘实体会构成一个有效的网络。根据上述描述，本文构建了数据科学招聘实体网络。

数据科学招聘实体网络的一个主要功能是能够提供数据科学职位的主要关注点，即通过数据科学招聘实体网络的节点，发现数据科学职位所共同关注的实体，而该实体主要是通过数据科学招聘实体网络中的中介度(centrality betweenness)获取的。

中介度(betweenness centrality)的概念最早用于分析社会网络中个体的重要性，由 L. C. Freeman[19] 在 1979 年提出。他认为，如果一个节点处于多对节点之间，该节点的度(degree)可能会较低。也就是说，如果只从度的角度来看，会误以为这个节点在网络中没有占据显著地位。但是，这个度较低的节点可能具有控制网络内部通信的重要作用，是网络中重要的节点。因此，中介度能够反映一个节点在网络中地位的重要程度，展现出其他节点对该节点的依赖程度。对于网络中一个节点 i，其中介度的计算公式为：

$$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \qquad \text{式(14)}$$

这里 $g_{jk}(i)$ 是节点对 j 和 k 之间，经过节点 i 的最短路径条数，g_{jk} 是连接节点 j 和 k 的所有最短路径的条数。那么 $g_{jk}(i)/g_{jk}$ 表示节点 j,k 之间经过 i 的最短路径条数占 j,k 间总的最短路径条数的比例。

中介度刻画了节点在网络中的重要程度，反映了节点控制网络内部通信的能力。一个节点的中介度越大，该节点的位置就越接近与整个网络中央，即该节点的地位越显著。正是由于该节点相对处于网络中心位置，使得网络中大量信息将要通过它，所以该节点对整个网络的内部通信控制具有重要的作用，节点本身也显得尤为重要。此外，中介度还可以反映整个网络的集中化程度。网络的集中化程度是检验复杂网络是否成熟的重要标志，如果整个网络的中介度较高，那么预示着该网络的成熟度(maturity)也达了到相对较高水平，整个网络呈现稳定和成熟的状态。

由于整个网络规模太大，难以全部展示，为了帮助理解，本文给出了基于数据科学语料库中 100 篇语料的两个小规模网络，并将编制的.net 格式文件导入 Pajek 软件绘制数据科学实体网络示例图。图 7 给出了由学历要求、专业要求、经验要求和能力要求构成的综合数据科学实体网络，图 8 给出了仅由软件实体构成的单一数据科学实体网络。

本文基于已经构建的数据科学实体网络，以软件实体为分析的样例，按照中介度降序分别筛选出了前 20 的数据科学软件实体，即数据科学招聘软件实体中的重要关注点，表 4 给出了前 20 个按中介度降序排列的数据科学软件实体。

表 4 中介度最高的 20 个数据科学软件实体

中介度排序	中介度	数据科学软件实体
1	0.1429695540	SQL
2	0.1305702550	Oracle
3	0.1045620190	MySQL
4	0.0843379040	Hadoop
5	0.0842922990	Java
6	0.0767000540	Excel
7	0.0643107090	C
8	0.0616286510	Python
9	0.0385161950	Linux
10	0.0338845110	Spark
11	0.0313447170	office
12	0.0275843660	R
13	0.0241636320	SAS
14	0.0240613100	SQLserver
15	0.0209335340	PPT
16	0.0200576790	IT
17	0.0193000400	BI
18	0.0162309970	ETL
19	0.0152836320	SPSS
20	0.0131819250	matlab

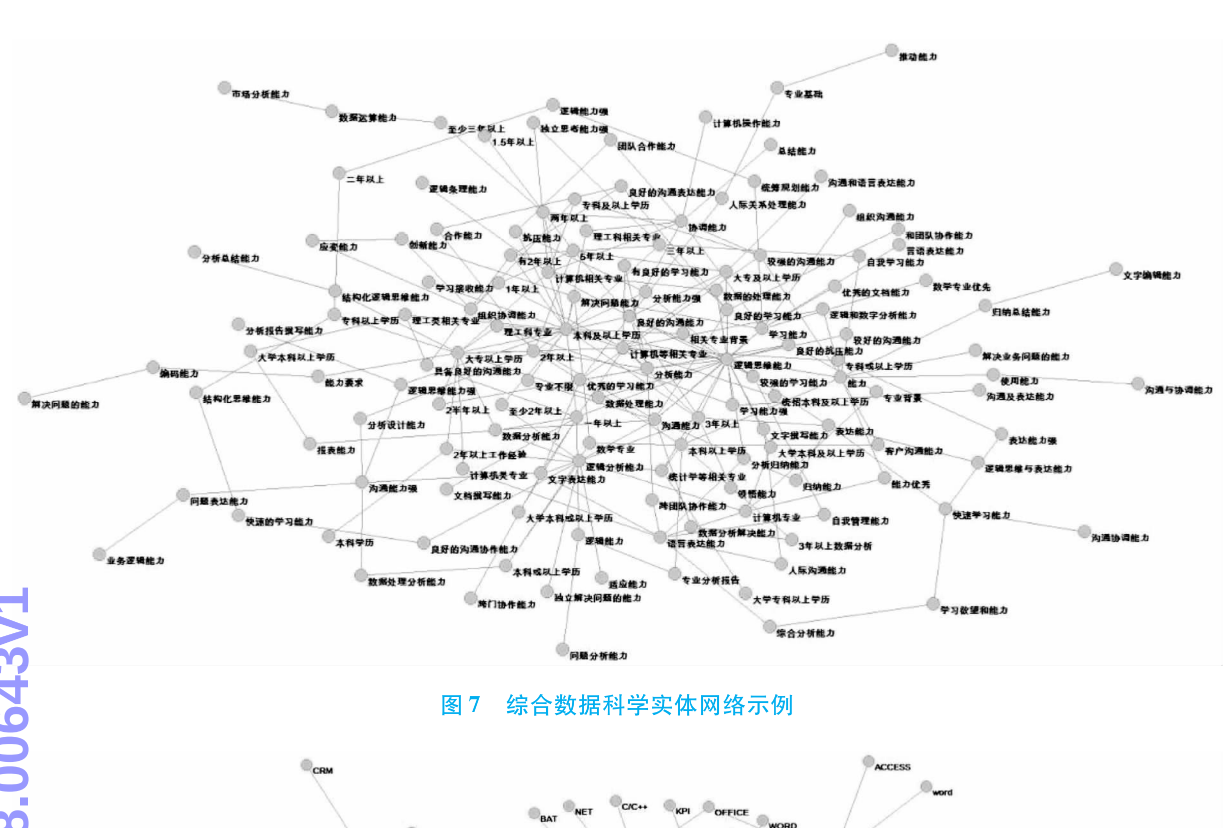


图 7　综合数据科学实体网络示例

图 8　单一数据科学实体网络示例

从表 4 可以看出，根据中介度值大于 0.1 进行筛选，得到排名前三的软件实体为“SQL”“Oracle”和“MySQL”。这 3 个实体要么是数据库要么是进行数据库操作的标准语句，是进行数据分析和挖掘的基础和前提，其排名非常靠前也充分说明了要进行数据分析或者挖掘首要必须完成对数据的存储和检索，而“SQL”位于第一位也充分说明了这一点。在最近三年内，大数据得到了快速发展，与大数据相关的技术在数据科学的招聘中体现得也比较充分，在前 20 的软件实体中，与大数据密切相关的技术就涉及到“Hadoop”“Spark”这两个实体，并且“Hadoop”在所有的软件实体当中排名位于第四。这一数据在一定程度上说明了数据科学与大数据之间存在着非常密切的关系，同时也说明了在后续的数据科学课程中要增加与大数据技术相关的教学内容。与编程语言相关的前 20 个实体主要涵盖了“Java”“C”“Python”“R”和“matlab”等 5 个，

而“Python”在数据科学招聘中是异军突起的一种编程语言,因为这一语言非常适合于处理数据尤其是非结构化数据,所以无论是在后续的课堂学习还是职业培训中,均应结合具体的数据处理任务强化对这一程序设计语言的教学。数据科学与统计学有着千丝万缕的联系,在一定程度上统计学支撑了数据科学的整个框架和体系,而在前 20 个实体中,“SPSS”和“SAS”这两个实体的入选也充分说明了这一点。虽然 office 是最基础的办公软件,但在数据的处理和呈现上有其独特之处,而“Excel”“PPT”和“office”等软件实体的入选,有力地证明了这一点,因为数据科学的职位中,不仅涉及到模型的构建、算法的设计这些相对技术难度比较大的职位,也涵盖了初级数据分析师和数据标注师这些技术难度一般但需求量较大的职位,而这些职位所使用的软件工具主要集中于“Excel”“PPT”等常用的软件工具上。受制于论文的篇幅,本文只给出了排名居于前 20 的软件实体在数据科学实体网络中的分布情况,并结合相应的招聘需求对典型的软件实体进行了分析。

6 结语

本文所研究的数据科学招聘实体自动抽取模型对于构建与数据科学实体相关的知识库和培养数据科学人才起到了充当基础资源的作用。本文在已标注的数据科学招聘实体的语料基础上,通过对比 Bi-LSTM-CRF、CRF 和 Bi-LSTM 这 3 个模型在实体招聘上的整体性能,不仅证明了深度学习模型在序列化识别任务上的优越性能,而且最终确定了由 Bi-LSTM-CRF 所构建的实体识别模型为数据科学招聘实体抽取的模型。并在这一模型的基础上,搭建了数据科招聘实体抽取平台和构建了基于 23 154 条数据科学招聘信息的实体网络,并对网络中的软件这一实体进行了分析。在后续的研究中,一方面要在各大招聘网站上使用该模型进行具体的应用推广,另一方面结合模型的整体性能表现,通过融合新的特征改进已有模型的精确率和召回率,从而提高数据科学招聘实体自动抽取模型的性能。

参考文献:

[1] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name [J]. Machine learning, 1999, 34 (1/3): 211 -231.

[2] BERGER A L, PIETRA V J D, PIETRA S A D. A maximum entropy approach to natural language processing [J]. Computational linguistics, 1996, 22(1): 39 -71.

[3] LAFFERTY J, MC CALLUM A, PRREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the eighteenth international conference on machine learning. San Francisco: Margan Kaufmann, 2001: 282 -289.

[4] MC CALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons [C]//Proceedings of the seventh conference on natural language learning at HLT-NAACL. Association for Computational Linguistics, 2003: 188 -191.

[5] 张小衡,王玲玲. 中文机构名的识别与分析[J]. 中文信息学报,1997,11(4): 21 -32.

[6] ZhANG Y, ZhOU J F. A trainable method for extracting Chinese entity names and their relations[C]// The Workshop on Chinese Language Processing: Held in Conjunction with the, Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2000:66 -72.

[7] 郑逢强,林磊,刘秉权.《知网》在命名实体识别中的应用研究[J]. 中文信息学报, 2008, 22(5): 97 -101.

[8] 陈宇,郑德权,赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报,2012,23(10):2572 -2585.

[9] 邵发,黄银阁,周兰江,等. 基于实体消歧的中文实体关系抽取[J]. 山东大学学报(工学版),2014,44(6):32 -37.

[10] 许华,刘茂福,姜丽,等. 基于语言规则的病症菌实体抽取[J]. 武汉大学学报(理学版),2015,61(2):51 -55.

[11] 冯蕴天, 张宏军, 郝文宁,等. 基于深度信念网络的命名实体识别[J]. 计算机科学, 2016, 43(4):224 -230.

[12] DONG C, ZHANG J, ZONG C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [C]// International conference on computer processing of oriental languages. New York City: Springer International Publishing, 2016:239 -250.

[13] 朱丹浩, 杨蕾, 王东波. 基于深度学习的中文机构名识别研究——一种汉字级别的循环神经网络方法[J]. 现代图书情报技术, 2016, 32(12):36 -43.

[14] 叶鹰, 马费成. 数据科学兴起及其与信息科学的关联[J]. 情报学报, 2015(6):575 -580.

[15] 杨京, 王效岳, 白如江,等. 大数据背景下数据科学分析工具现状及发展趋势[J]. 情报理论与实践, 2015, 38(3):134 -137.

[16] 周傲英, 钱卫宁, 王长波. 数据科学与工程:大数据时代的新兴交叉学科[J]. 大数据, 2015, 1(2):90 -99.

[17] 朝乐门, 卢小宾. 数据科学及其对信息科学的影响[J]. 情报学报, 2017, 36(8):761 -771.

[18] 王曰芬, 谢清楠, 宋小康. 国外数据科学研究的回顾与展望[J]. 图书情报工作, 2016, 60(14):5 -14.

[19] FREEMAN L C. Centrality in social networks conceptual clarification[J]. Social networks, 1979,1 (3),215 -239.

王东波,胡昊天,周鑫,等.基于深度学习的数据科学招聘实体自动抽取及分析研究[J].图书情报工作,2018,62(13):64-73.

作者贡献说明:

王东波:提出论文框架,设计算法并撰写论文;

胡昊天:数据标注与模型参数调整和论文撰写;

周鑫:测试性能评价;

朱丹浩:模型训练。

Research of Automatic Extraction of Entities of Data Science Recruitment and Analysis Based on Deep Learning

Wang Dongbo[1] Hu Haotian[1] Zhou Xin[2] Zhu Danhao[3]

[1] College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095

[2] Department of Information Management, Nanjing University, Nanjing 210093

[3] Department of Computer Science and Technology, Nanjing University, Nanjing 210093

Abstract: [**Purpose/significance**] Data science is emerging as a new interdisciplinary field which combines many fields. Extracting the corresponding entities knowledge from the announcement information of data science recruitment can not only help to understand the development of data science from a market perspective, but also help to improve the content of data science teaching. [**Method/process**] Based on the recruitment announcement from the recruitment website, combining with information science data collection, annotation and organization methods, data science corpus was constructed and the corresponding entities from it were extracted. [**Result/conclusion**] In the existing 11000 annotated data science corpus scale recruitment announcement, based on the Bi-LSTM-CRF, CRF and Bi-LSTM models, this paper compared the extraction performance of data science recruiting entities and finally determined the final data science recruitment entities automatic extraction model, designed the data science recruitment entities automatic extraction platform, and built a data science recruitment entities network.

Keywords: data science　conditional random field　deep learning　Bi-LSTM-CRF